

# WTF? Discovering the Unexpected in next-generation radio continuum surveys

Evan Crawford<sup>1</sup>, Ray P. Norris<sup>1,2</sup>, and Kai Polsterer<sup>3</sup>

<sup>1</sup>*Western Sydney University, Penrith, NSW, Australia;*  
e.crawford@westernsydney.edu.au

<sup>2</sup>*CSIRO Astronomy & Space Science, Epping, NSW, Australia*  
raypnorris@gmail.com

<sup>3</sup>*Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*  
Kai.Polsterer@h-its.org

**Abstract.** Most major discoveries in astronomy have come from unplanned discoveries made by surveying the Universe in a new way, rather than by testing a hypothesis or conducting an investigation with planned outcomes. Next generation radio continuum surveys such as the Evolutionary Map of the Universe (EMU: the radio continuum survey on the new Australian SKA Pathfinder telescope), will significantly expand the volume of observational phase space, so we can be reasonably confident that we will stumble across unexpected new phenomena or new types of object. However, the complexity of the instrument and the large data volumes mean that it may be non-trivial to identify them. On the other hand, if we don't, then we may be missing out on the most exciting science results from EMU. We have therefore started a project called “WTF”, which explicitly aims to mine EMU data to discover unexpected science that is not part of our primary science goals, using a variety of machine-learning techniques and algorithms. Although targeted specifically at EMU, we expect this approach will have broad applicability to astronomical survey data.

## 1. Introduction

Major scientific discoveries, such as the discovery of the Higgs Boson (Aad et al. 2012), are often made by testing a hypothesis or conducting an investigation with planned outcomes. However, most major discoveries in astronomy are unplanned and unexpected (Harwit 1981; Wilkinson et al. 2004; Ekers 2009; Norris et al. 2015; Norris 2015). Typically these occur as the result of building larger telescopes, or opening up a new window of the electromagnetic spectrum. More generally, we may define a phase space whose axes correspond to observable quantities. Some parts of this phase space (e.g. bottom right of Fig. 1) have been well-observed and have already yielded their discoveries, whereas some parts of this space (e.g. top left of Fig. 1) have not yet been observed, and may contain new discoveries that are available to new instruments sampling that region of phase space. Virtually all “accidental” or “serendipitous” discoveries result from observing a new part of this phase space (Norris 2015). The case study of the Nobel-prize-winning discovery of pulsars by Jocelyn Bell is instructive. A talented and persistent PhD student studying interstellar scintillation (and thus expanding the obser-



## 2. Widefield ouTlier Finder

Widefield ouTlier Finder (WTF) is a pilot project to test and implement techniques for mining large data sets for unexpected discoveries. WTF takes the overall approach initially of constructing large data sets containing simulated discoveries (named “eggs”) and then inviting competing algorithms and techniques to find these eggs in a series of ‘data challenges’. Visualisation tools will also be developed to aid understanding of the process and its results. The data include both real and simulated data, and include both images (which may include spectral shape and polarisation data) and tabular data, and will include the multiwavelength identifications and other properties where available.

Different algorithms are then applied in blind tests to see which are most successful at finding eggs. This process includes inviting external collaborators to try their algorithms on the data. Because some of the data is real data that has not been mined, there is a chance, even at this early stage, that the teams may make genuine unexpected discoveries. In later stages of the project, the most successful techniques will be used to search for discoveries in real EMU data. The large data volumes necessitate automatic processing, with minimal human interaction, providing a challenge well-suited to machine learning techniques.

We are implementing WTF initially as a collaborative environment on the Amazon Web Services<sup>1</sup> platform, which provides disk space, processing capacity, and infrastructure. The AWS platform is well-suited to this project, in providing a collaborative research environment, with disk and processing capacity that can be varied to suit the demand.

We expect many different algorithms to be applied to the task of finding the eggs. For example, one approach may look for groups of properties in an  $n$ -dimensional plot with axes such as flux density, spectral index, and IR-to-radio ratio. Some groups will correspond to known types of object (e.g. stars, galaxies, quasars) but unexpected groups may correspond to unknown classes. More sophisticated approaches will use machine-learning algorithms such as neural nets, dimensionality reduction (self-organising maps, autoencoding, isoplanar mapping, TSNE, etc.), clustering (k-means, db-scan, Birch), outlier detection, and Bayesian approaches. Although targeted at EMU, such approaches will be widely applicable to astronomical survey data.

It is likely that different algorithms will be most suited for particular challenges, and so, even in the latest stages of this project, when EMU data is mined for unexpected discoveries, a variety of different algorithms will still be used. Furthermore, it is likely that new algorithms and new approaches will evolve during the lifetime of this project, so a cyclic process is envisaged where the data challenges are followed by periods of algorithm development followed by further challenges.

We intend to run the WTF project in 4 phases, as follows. Only the first phase is currently funded.

- Phase 1: Set up infrastructure, and initial challenges consisting of data (images or tables) with embedded “eggs”; apply the first challenger algorithms to debug and refine the infrastructure. Validate the architecture using in-house challengers; test that the infrastructure is extendable to a larger number of users.

---

<sup>1</sup><http://aws.amazon.com>

- Phase 2: Test different approaches and algorithms, to see which are best at discovering WTFs, and refine challenges; perhaps even make a real discovery!
- Phase 3: Mount ASKAP early science challenge, probably find artefacts (which will aid in ASKAP commissioning), test different approaches and algorithms, to see which are best at discovering WTFs in REAL ASKAP data; increased probability of making a real discovery!
- Phase 4: By 2017 the EMU survey should be well underway, so we can start adding the EMU data to the WTF machinery. At this stage there is an increasing likelihood of making a real discovery.

### 3. Conclusion

WTF is the first stage of our long-term goal of an extensive international collaboration to apply machine learning and other data science techniques to mine large survey data for new scientific knowledge.

Our early goals include: (a) building infrastructure and techniques for mining the unexpected from astronomical data, and (b) developing experience in using cloud computing as a collaborative research environment. We also plan to develop, with our collaborators, a set of best practice guidelines for:

- inviting collaborators to a project like this, including expectation management
- allocation of resources to collaborators
- monitoring the usage of and tuning allocations
- a publishing model for results generated from this work

### Acknowledgements

We thank Amazon Web Services (AWS) for providing a grant to kick-start WTF. We thank all members of the WTF team for contributing to this project.

### References

- Aad, G., et al. 2012, Physics Letters B, 716, 1
- Bell Burnell, J. 2009, in Accelerating the Rate of Astronomical Discovery, [pos.sissa.it/cgi-bin/reader/conf.cgi?confid=99,14](http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=99,14)
- Ekers, R. D. 2009, in Accelerating the Rate of Astronomical Discovery, [pos.sissa.it/cgi-bin/reader/conf.cgi?confid=99,7](http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=99,7)
- Harwit, M. 1981, Cosmic Discovery (MIT press)
- Johnston, S., et al. 2008, Experimental Astronomy, 22, 151. 0810.5187
- Norris, R. 2015. In preparation
- Norris, R., Basu, K., Brown, M., Carretti, E., Kapinska, A. D., Prandoni, I., Rudnick, L., & Seymour, N. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 86. 1412.6076
- Norris, R. P., et al. 2011, PASA, 28, 215. 1106.3219
- Wilkinson, P. N., Kellermann, K. I., Ekers, R. D., Cordes, J. M., & W. Lazio, T. J. 2004, New A Rev., 48, 1551. astro-ph/0410225